A Semi-supervised Word Alignment Algorithm with Partial Word Alignment

Qin Gao, Nguyen Bach and Stephan Vogel
Language Technologies Institute
Carnegie Mellon University
# Supervised and unsupervised word alignment

## Unsupervised word alignment:
- GIZA++ (IBM Models)
- LEAF (Fraser and Marcu, 2007)

## Supervised word alignment:
- Maximum Weight Bipartite Matching (Taskar et al, 2005; Lacoste-Julien et al, 2006)
- Maximum Entropy (Ittycheriah and Roukos, 2005)

## Semi-supervised word alignment

### Discriminative models:
- Blunsom and Cohn (2006)
- Niehues and Vogel (2008)
- Taskar et al (2005)
- Moore (2005)

### Discriminative training of generative models:
- Model 6 (Och and Ney, 2003)
- Fraser and Marcu (2006)

### Mixing parameters trained from labeled and unlabeled data:
Incomplete reference?

伊犁 大规模 开展 " 面对面 " 宣讲 活动

Xinjiang ‘s Yili holds propaganda drive

An oracle that only has partial knowledge
E.g. Location name dictionary, technical term dictionary
What is the difference of partial and full word alignment?

伊犁大规模开展“面对面”宣讲活动

Xinjiang’s Yili holds propaganda drive

伊犁大规模开展“面对面”宣讲活动

Xinjiang’s Yili holds propaganda drive

AER? Entropy? Likelihood?
How to make use of the partial knowledge?

Most straightforward way: Post-process
A better way

• Let the knowledge determine the known part, and let models determine the rest.

• The knowledge will:
  – Affect the statistics we get for the model
  – Be reflected in the final alignment

Anything conflicting with known alignments should be forbidden

• Pereira and Schabes, 1992, Similar idea on SCFGs
A summary of the proposed method

EM Algorithm

Constraints from partial alignments

Make sure the alignments that are inconsistent with the constraints get zero probability
Recap: Training of IBM Models

EM Algorithm

Model 1,2: Maximizing $L(F,A|E,\Theta)$ can be done directly over all possible alignments

Model 3,4,5: Even finding Viterbi alignment is too expensive

Generate seed alignment from HMM/Model 2

Find local optimal “center alignment” using Hill-climbing

Collect statistics from the neighbors of the center alignment
Hill-climbing

Initial "center alignment"

Are there better alignment in the "neighborhood"

Change the center alignment to the best alignment in the neighborhood

Output the center alignment as the "Viterbi" alignment
Neighborhood of an alignment

Move operator $M_{1,1}$

Swap operators $S_{7,8}$

All the alignments that are different from the current center alignment by \textit{ONE} operator, are considered in the neighborhood of it.
Applying the idea on fertility-based models

- Ruling out the inconsistent alignments during the hill-climbing stage
  - Yield a consistent, yet optimal “Viterbi” alignment
- Ruling out the statistics of inconsistent alignments during the statistics collection stage.
  - The probability of consistent alignments will increase in the next iteration
Two types of constraints

Target sentence: 伊犁 大规模 开展 " 面对面 " 宣讲 活动
Source sentence: Xinjiang’s Yili holds propaganda drive

Type 1: A Target word aligned to a different source word
Type 2: A Source word is aligned to a concrete target word, but we know it should not be aligned to any.
Modified Hill-climbing

Every iteration, pick the alignment:
1. Corrects at least one inconsistent link
2. Has highest probability

Every iteration, pick the alignment:
1. Does not introduce inconsistent links
2. Has highest probability

Xinjiang’s Yili holds propaganda drive

Correct the inconsistent links
Optimize towards local optimal
Statistics collection

• Navigate through all the neighbor alignments and only collect statistics of consistent alignments.

• The modified hill-climbing and statistics collection can be done efficiently by manipulating the moving/swapping matrices.
A summary of the proposed method

EM Algorithm

Constraints from partial alignments

Make sure the alignments that are inconsistent with the constraints get zero probability

Modified Hill-climbing

Statistics collection of consistent alignments
Experiment Design: What’s the behavior of the method?

- Does the algorithm correct more links than simple post processing?
- Does having the answer for different part of the sentence makes difference?
- Is the method robust to the sub-optimal oracles (aka, error-prone answers)?

- We can use the method to exploit precious partial manual alignments.
- We can use active learning methods to choose the most important words to label.
- We can integrate the method with other automatic word aligners.
Experiment 1: does it correct more than what we already know?

- The proposed method vs. post processing
- Randomly select different numbers alignment links
- Corpus: Chinese-English and Arabic-English manually aligned corpus

<table>
<thead>
<tr>
<th></th>
<th>Sentence Pairs</th>
<th>Num. of Words</th>
<th>Num. of Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese-English</td>
<td>21K</td>
<td>424K/524K</td>
<td>687K</td>
</tr>
<tr>
<td>Arabic-English</td>
<td>29K</td>
<td>630K/821K</td>
<td>830K</td>
</tr>
</tbody>
</table>
Alignment Error Rate Ch-En
Random Selection

AER

Number of used alignment links (x1000)

Post-processing
Alignment Error Rate Ch-En
Random Selection

Number of used alignment links (x1000)

AER

0  100  200  300  400  500  600  700

Semi-Supervised Aligner
Post-processing
Alignment Error Rate Ch-En
Random Selection

![Graph showing Alignment Error Rate (AER) vs. Number of used alignment links (x1000). The graph compares Semi-Supervised Aligner, Post-processing, and Difference.](image)
Alignment Error Rate Ar-En
Random Selection

Number of used alignment links (x1000)

AER

Semi-Supervised Aligner
Post-processing
Difference
Experiment 2: Carefully selecting alignment links makes a difference

• We try to find a better way of selecting alignment links from manual alignment

• Intuitively: Select alignment links that two GIZA++ directions disagree about.
Alignment Error Rate Ch-En
Random Selection v.s Disagreement-based

![Graph showing the comparison between Random Selection and Disagreement-based Selection for the Alignment Error Rate (AER)]

- **Random Selection**
- **Disagreement-Based Selection**
- **Difference**

**Axes:**
- Y-axis: AER
- X-axis: Number of used alignment links (x1000)

**Legend:**
- Red: Random Selection
- Green: Disagreement-Based Selection
- Blue Dashed: Difference

**Data Points:**
- At 0 alignment links, AER is around 42 for Random Selection and 37 for Disagreement-Based Selection.
- As the number of alignment links increases, the AER decreases for both methods, with the Disagreement-Based Selection consistently lower than Random Selection.
- The difference between the two methods becomes more pronounced as the number of alignment links increases.
Alignment Error Rate Ar-En
Random Selection v.s Disagreement-based

![Graph showing the comparison between Random Selection and Disagreement-based Selection for Alignment Error Rate (AER) with respect to the number of used alignment links (x1000). The graph illustrates the decreasing trend in AER as the number of alignment links increases, with Random Selection and Disagreement-based Selection showing distinct curves. The difference curve also highlights the disparity between the two methods.](image-url)
Experiment 3: Does it works for simple (AKA error-prone) heuristics?

• We used LDC Chinese-English dictionary to generate high-precision-low-recall partial alignments

• The entries with single Chinese character or more than six English words are filtered out.

• Add links when a lexicon entry was encountered in the sentence pair

• 79.48% precision and 17.36% recall rate
Alignment quality

<table>
<thead>
<tr>
<th></th>
<th>Chinese-English</th>
<th>English-Chinese</th>
<th>grow-diag-final-and</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>68.22 (+1.71)</td>
<td>65.35 (+1.35)</td>
<td>69.15 (+0.96)</td>
</tr>
<tr>
<td></td>
<td>69.93</td>
<td>66.70</td>
<td>70.11</td>
</tr>
<tr>
<td>Recall</td>
<td>46.88 (+1.40)</td>
<td>55.05 (+1.40)</td>
<td>57.47 (+2.07)</td>
</tr>
<tr>
<td></td>
<td>48.28</td>
<td>56.45</td>
<td>59.54</td>
</tr>
<tr>
<td>AER</td>
<td>44.43 (-1.43)</td>
<td>40.28 (-1.55)</td>
<td>37.23 (-1.62)</td>
</tr>
<tr>
<td></td>
<td>42.88</td>
<td>38.85</td>
<td>35.61</td>
</tr>
</tbody>
</table>

10/13/2010
The improvement on translation qualities (400k sentence pairs)

<table>
<thead>
<tr>
<th></th>
<th>MT02</th>
<th>MT03</th>
<th>MT04</th>
<th>MT05</th>
<th>MT08NW</th>
<th>MT08WB</th>
<th>Dev07NW</th>
<th>Dev07WB</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>28.87</td>
<td>27.82</td>
<td>30.08</td>
<td>26.77</td>
<td>25.09</td>
<td>17.72</td>
<td>24.88</td>
<td>21.76</td>
<td>25.37</td>
</tr>
<tr>
<td>Dict-Link</td>
<td>29.59</td>
<td>27.67</td>
<td>31.01</td>
<td>27.13</td>
<td>25.14</td>
<td>17.96</td>
<td>25.51</td>
<td>21.88</td>
<td>25.74</td>
</tr>
</tbody>
</table>

BLEU Scores
Conclusion

• We implemented a semi-supervised word alignment algorithm based on IBM models which can use partial word alignment.

• Experiments were performed to prove that:
  – 1. The algorithm can correct more links than directly fixing the incorrect links
  – 2. Better alignment quality can be achieved by carefully selecting words to ask the oracle
  – 3. By supplying high-precision-low-recall alignment links the alignment quality can also be improved.
Applications

• Integrate manual alignments
  – Use Mechanical Turk to collect partial alignments
  – Reference: Gao and Vogel, MTurk workshop 2010

• Active learning for word alignment
  – Select the most significant words to ask the oracles for answers
  – Reference: Ambati, Vogel and Carbonell, ACL 2010

• Interact with supervised aligners
  – Use supervised aligner to generate high-precision-low-recall alignment links
  – Iteratively improve both aligners
  – Reference: Gao, Guzman and Vogel, COLING 2010
THANK YOU
The improvement on alignment qualities

<table>
<thead>
<tr>
<th></th>
<th>Chinese-English</th>
<th>English-Chinese</th>
<th>grow-diag-final-and</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>AER</td>
</tr>
<tr>
<td>Baseline</td>
<td>68.22</td>
<td>46.88</td>
<td>44.43</td>
</tr>
<tr>
<td>Dict-Link</td>
<td>69.93</td>
<td>48.28</td>
<td>42.88</td>
</tr>
</tbody>
</table>
Priority of alignment links

• Run GIZA++ on both directions
• Pick the words that have different alignments in two directions and ask for answers from the oracle

- $C1$: $f_j$ aligns to $e_i, i > 0$ in $e \rightarrow f,^1$ but in reversed direction $e_i$ does not align to $f_j$ but to another word.
- $C2$: $f_j$ aligns to $e_i, i > 0$, in $f \rightarrow e$, but in reversed direction ($e \rightarrow f$), $f_j$ aligns to the empty word.
- $C3$: no word aligns to $f_j$, in $f \rightarrow e$, but in reversed direction $f_j$ aligns to $e_i, i > 0.\!^2$

Table 1: The priorities of alignment links

<table>
<thead>
<tr>
<th>Order</th>
<th>Criterion</th>
<th>Order</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$f_j \in C1$</td>
<td>5</td>
<td>$e_i \in C2$</td>
</tr>
<tr>
<td>2</td>
<td>$f_j \in C2$</td>
<td>4</td>
<td>$e_i \in C1$</td>
</tr>
<tr>
<td>3</td>
<td>$f_j \in C3$</td>
<td>6</td>
<td>$e_i \in C3$</td>
</tr>
</tbody>
</table>
The result by applying our method

Fix the known alignment links during training, instead of fixing it afterwards.

By fixing the link between Yili and 伊犁, other links also get corrected. (Of course, new errors may be introduced)
Applications

• Using active learning method to collect useful alignment links from users
• Make use of noisy manual alignments from non-experts such as Mechanical Turk
• Integrate other word aligners e.g. discriminative word aligner
In the context of word alignment

\[ P(f_1^J | e_1^I) = \sum_{a_1^J} Pr(f_1^J, a_1^J | e_1^J, \theta) \]

\[ Pr(f_1^J | e_1^I, a_1^J, \alpha_1^I) = \begin{cases} 0, & \text{if } a_1^J \text{ is inconsistent with } \alpha_1^I \\ Pr(f_1^J | e_1^I, a_1^J, \theta), & \text{otherwise} \end{cases} \]
Overview

• A word alignment algorithm that can take partial alignment links and improve the unsupervised word alignment results

• The method enables integrating knowledge from external sources into unsupervised word alignment
In this talk

- The detail of the algorithm
- Integrating (oracle) manual alignments
- Applying alignment links from lexicon

Easy things easy (corrected)
Hard things possible (improved)
What is partial word alignment?

伊犁大规模开展"面对面"宣讲活动

Xinjiang‘s Yili holds propaganda drive
Modified Hill Climbing

• In Hill-climbing step 1: Pick the operator that
  1. eliminate at least one inconsistent link
  2. yield best probability
Modified Hill Climbing

• In Hill-climbing step 2:
  – Make a cell negative if the operator will bring in inconsistent link
  – Recalculate the signs of the cells in swap matrix if a swap operator is chosen
The M-Step

• In statistics collection:
  – Ignore the statistics from cells with negative values